Search of large-scale deletions in target-sequenced NGS data

German Demidov, scientific advisor: Anton Bragin

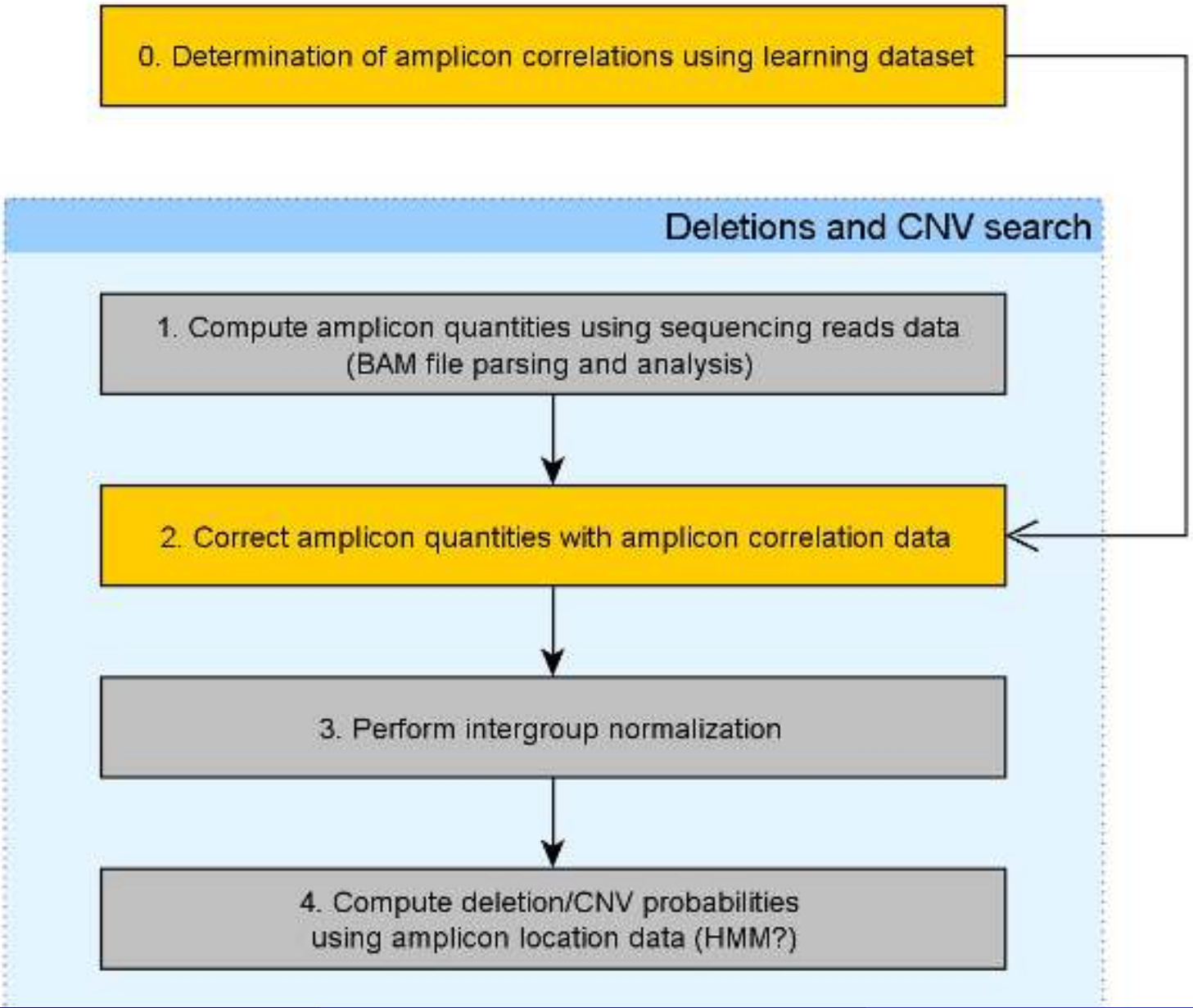
St.Petersburg University of the Russian Academy of Sciences

December 19, 2013

Problem

1. Problem definition: we need to know, if large-scale deletion happened or not and where and is a patient homozygote or heterozygote by this defective allele.
2. This is a critical step for clinical recognition of genetic hereditary diseases, such as cystic fibrosis.
3. Our data in the first step: simple .xml files with information of amplicons' coverage that have been obtained by using IonTorrent target sequencing of the different regions of CFTR-gene that located on 7th chromosome, PAH-gene located on 12th cromosome and GALT-gene located on 9th chromosome (connected with cystic fibrosis, phenylketonuria and galactosemia, accordingly).

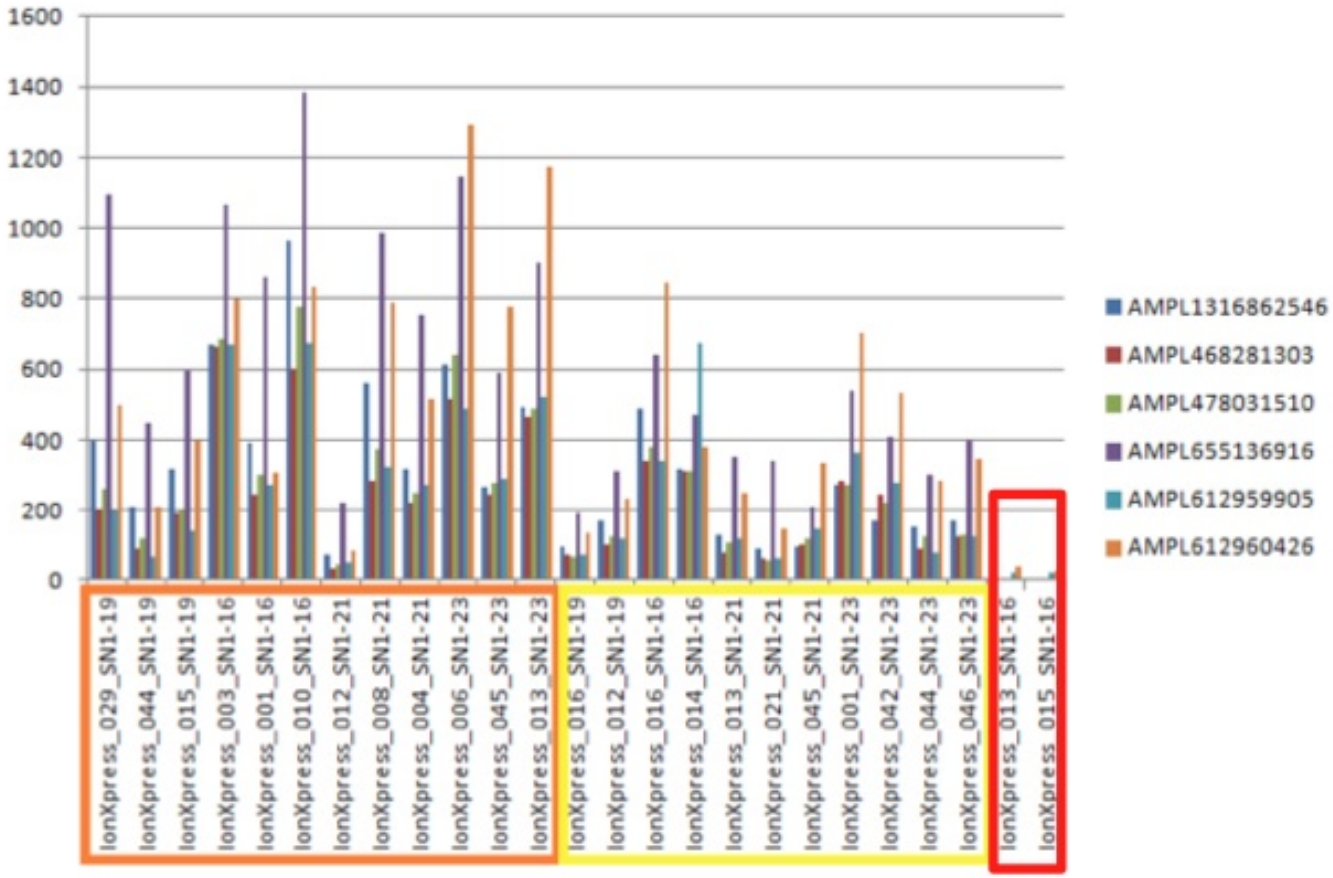
What we want to do.



Coverage-based analysis

1. Hence, we can look for the zero coverage of amplicon to understand, if this sample homozygote for the deletion in CFTR gene - gene, that is connected with cystic fibrosis. It is not so complicated - we can just build graph and look at it.
2. But we are interested in the diagnosis of deletion in heterozygote. It is not so simple. Coverage of deleterious region in chromosome is not equal zero if our patient is heterozygote.

There is indeterminacy of coverage



The main reason of this indeterminacy is the complicated biochemical mechanism of PCR.

LR.

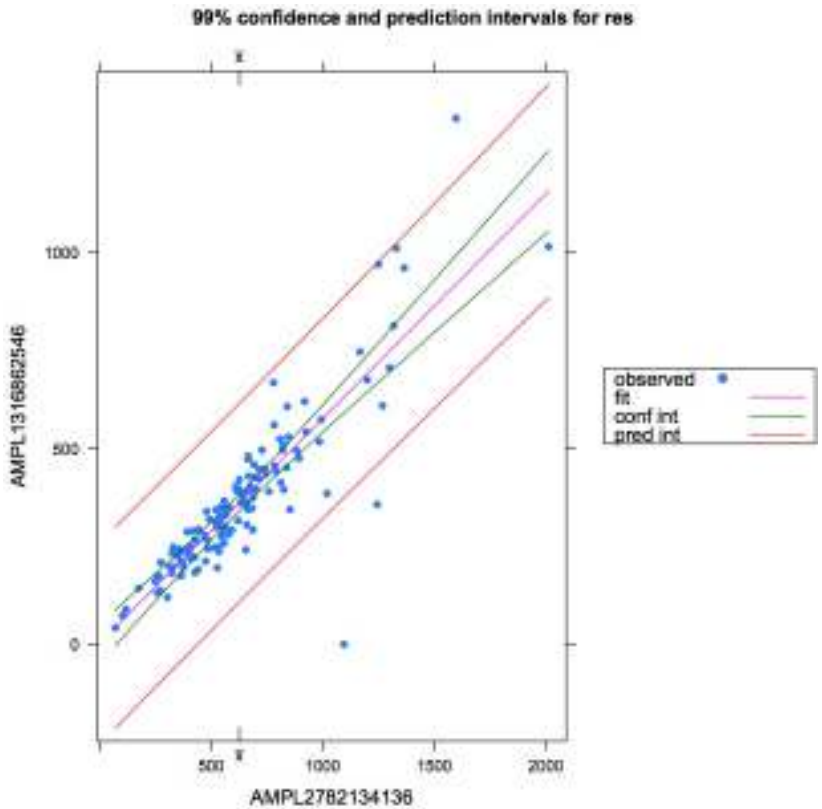


Figure: Look at this beautiful linear regression model.

Pay attention to prediction and confidence intervals. Can you see this dots below regression line? This dots mean structural variants, this samples are heterozygote/homozygote by deletion. And we want to detect this outliers.

Our data.

1. We had a set of 160 samples obtained in 4 experiments (12 of our samples was heterozygote, 3 of them was homozygote), each of samples consists of 129 amplicons, also we had different methods of clusterization, different regression models, normalizations and 2 books about statistics...and also the whole Internet to learn.
2. Not that we needed all that for the trip, but once you get locked into a serious collection, the tendency is to push it as far as you can.

More sophisticated normalization.

Then we want to increase the accuracy of our predictions. Hence we need to cluster our amplicons in several groups. How we can do it? At the first step, we decided to do it with the help of correlational analysis.

	AMPL1350926839	AMPL1350926838	AMPL1350926835	AMPL1425598476	AMPL1350926836	AMPL612959905	AMPL46
AMPL1350926839	1.0000000	0.9137302	0.95348572	0.9118448	0.9833118	0.93894211	0.9
AMPL1350926838	0.9137302	1.0000000	0.80670186	0.7105979	0.9246608	0.76429588	0.9
AMPL1350926835	0.9534857	0.8067019	1.00000000	0.9300787	0.9129884	0.96721786	0.8
AMPL1425598476	0.9118448	0.7105979	0.93007870	1.0000000	0.8768594	0.91702842	0.7
AMPL1350926836	0.9833118	0.9246608	0.91298844	0.8768594	1.0000000	0.92667961	0.9
AMPL612959905	0.9389421	0.7642959	0.96721786	0.9170284	0.9266796	1.00000000	0.8
AMPL468748906	0.9011938	0.9079783	0.81772612	0.7173869	0.9085780	0.82921164	1.0
AMPL474961767	0.7862506	0.5834210	0.86902134	0.7991670	0.6893262	0.82557678	0.6
AMPL1486260392	0.9379665	0.8358004	0.93016713	0.8168823	0.9199303	0.93403193	0.9
AMPL468865894	0.9477317	0.9138056	0.86350542	0.8059069	0.9554640	0.88364997	0.9
AMPL469688358	0.9640895	0.9218195	0.89138374	0.8332860	0.9840521	0.90194044	0.8
AMPL1489306062	0.9706267	0.9328851	0.93899210	0.8555511	0.9716858	0.91247953	0.8
AMPL469322657	0.9607553	0.9474740	0.88140289	0.7990236	0.9715162	0.88327631	0.9
AMPL612960517	0.9783916	0.8464488	0.96793919	0.9220345	0.9670230	0.97970070	0.8
AMPL663779199	0.9317480	0.9031943	0.88811266	0.8408733	0.9566834	0.87631571	0.8
AMPL556424352	0.9529712	0.8629218	0.95599648	0.8959556	0.9551192	0.94310271	0.8
AMPL623737664	0.8982256	0.9725335	0.76259161	0.7092468	0.9202229	0.73155136	0.8
AMPL467632467	0.8177194	0.5926388	0.89709192	0.9231440	0.7483317	0.84460698	0.5
AMPL467561353	0.9419456	0.8396823	0.95047495	0.9026116	0.9348789	0.94479007	0.8
AMPL469294494	0.9038501	0.7389954	0.94866913	0.8725791	0.8679138	0.95994107	0.8
AMPL467649362	0.4132771	0.2081274	0.60052209	0.4780865	0.2757166	0.51686560	0.3
AMPL468184487	0.4028730	0.3556688	0.42120459	0.2469204	0.3453550	0.38451093	0.5
AMPL468425396	0.9556974	0.8866416	0.95353994	0.8534116	0.9624795	0.95642020	0.8
AMPL807033687	0.9581423	0.9230366	0.90654803	0.8251770	0.9771376	0.90231709	0.8
AMPL556370172	0.8975759	0.8583309	0.89989778	0.8230638	0.9068768	0.87061780	0.7
AMPL467692120	0.9849758	0.9388893	0.91795918	0.8591705	0.9967734	0.92487743	0.9
AMPL623760758	0.8390560	0.9745036	0.70467649	0.5985264	0.8609529	0.65698659	0.8
AMPL655136916	0.9245529	0.9091628	0.87392684	0.7992123	0.8835426	0.81795086	0.7
AMPL712350278	0.9590019	0.7878408	0.93879930	0.9729838	0.9195573	0.92345625	0.7

More sophisticated normalization.

1. We can choose one of three methods of correlation definition: Pearson, Spearman and Kendall. We decided to use Pearson correlation because we wanted to use linear regression.
2. And this type of clusterization works well for our samples.
3. Our first resluts was not as bad as they could be. We have found a small number of false-positives, but we have also detected almost all deleterious regions. Not every amplicon, but a lot of them in each heterozygote sample.
4. Then we included “clean samples” in our learning sample. It is impossible to use only samples from the set that is under test in new learning sample - 'cos all samples from our set can be deleterious!

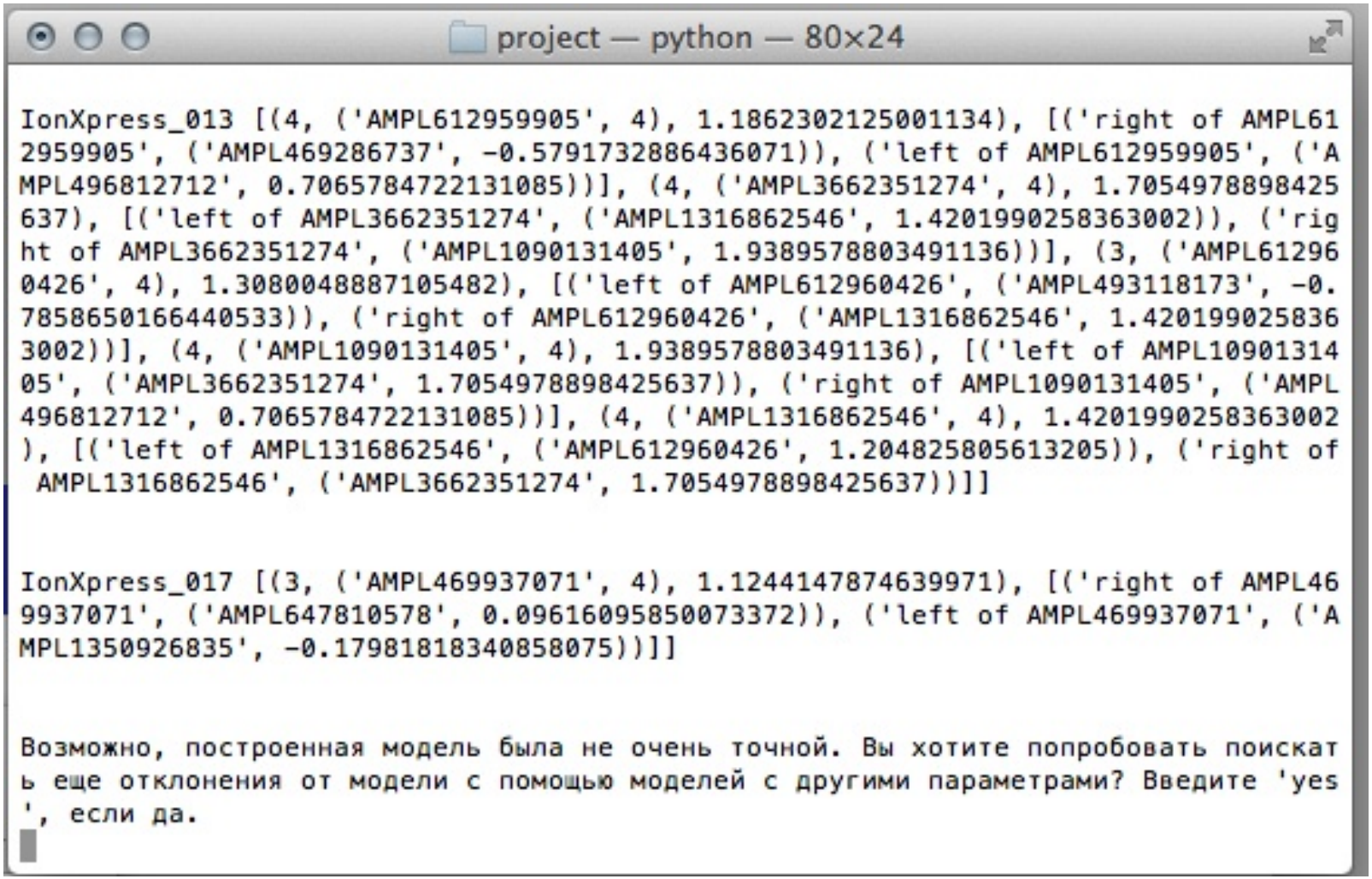
LR.

So, we used supervised learning.
We looked at neighboors of deleterious amplicon, we took deviations into account. Also we can estimate probabitities, errors, coefficient of determination, but we do not know how to analyse this set of numbers.
Also we have tried different methods to determine a deviation of amplicon's coverage, and have decided to use measure of deviations of amplicon with coverage x_0 (second amplicon from our cluster had coverage y_0, n means number of samples in our set, q - t-student distribution quantile):

$$\frac{(\hat{a} + \hat{b} \cdot x_0 - y_0)}{q \cdot \sigma^2 \cdot \left(1 + \frac{1}{n} + \sqrt{\frac{(x_0 - E(X))^2}{(n-1) \cdot (\sigma(X))^2}}\right)}$$

We rotated and moved our regression model because we wanted to estimate an error of our linear regression by calculation confidence intervals for parameters of linear regression.

Result.


project — python — 80×24
IonXpress_013 [(4, ('AMPL612959905', 4), 1.1862302125001134), [('right of AMPL612959905', ('AMPL469286737', -0.5791732886436071)), ('left of AMPL612959905', ('AMPL496812712', 0.7065784722131085))], (4, ('AMPL3662351274', 4), 1.7054978898425637), [('left of AMPL3662351274', ('AMPL1316862546', 1.4201990258363002)), ('right of AMPL3662351274', ('AMPL1090131405', 1.9389578803491136))], (3, ('AMPL612960426', 4), 1.3080048887105482), [('left of AMPL612960426', ('AMPL493118173', -0.7858650166440533)), ('right of AMPL612960426', ('AMPL1316862546', 1.4201990258363002))], (4, ('AMPL1090131405', 4), 1.9389578803491136), [('left of AMPL1090131405', ('AMPL3662351274', 1.7054978898425637)), ('right of AMPL1090131405', ('AMPL496812712', 0.7065784722131085))], (4, ('AMPL1316862546', 4), 1.4201990258363002), [('left of AMPL1316862546', ('AMPL612960426', 1.204825805613205)), ('right of AMPL1316862546', ('AMPL3662351274', 1.7054978898425637))]]
IonXpress_017 [(3, ('AMPL469937071', 4), 1.1244147874639971), [('right of AMPL469937071', ('AMPL647810578', 0.09616095850073372)), ('left of AMPL469937071', ('AMPL1350926835', -0.17981818340858075))]]
Возможно, построенная модель была не очень точной. Вы хотите попробовать поискать еще отклонения от модели с помощью моделей с другими параметрами? Введите 'yes', если да.

Plans.

1. How we can decide what level of confidence we should choose? Some experiments gave very clean results, so we needed to decrease the level of confidence to use. Today we decided to use 90% quantile of t-student distribution, but we want to measure total disperision of all samples got from experiment. How? Dunno. But we will do it.

Plans.

1. And when we want to add “clean samples” from the set under test - how we can determine them? Now we use simple measures and magic constants, and it works. May be it should be more sophisticated? Yes, we want to determine using some kind of mathematical function.

Plans.

1. Now we can look at the closest neighboors of deleterious amplicon. But we want to make a graphical map of chromosome with structural variant! How to do this? Easy and simple. We just need more time.

Plans.

1. Finally we can vary parameters of our linear regression model. But all we have - just a new list of deleterious amplicons. We want to make a rigorous definition of new lists!

Questions and answers?

Figure: Thank you!